CAMDA Forensics Challenge: An Evaluation of Mass-Transit, Microbiome Profiles

Keenan Berry, Jason Holdener, Yu Zhan

Abstract

Microbial communities exist in virtually every environment found within our planet, including subway transit systems. As a challenge to the global bioinformatics community, the MetaSUB International Consortium, has developed a challenge of classifying the location(s) of genetic samples obtained from various surfaces found in subway systems of 16 global cities. All samples used in our study were provided by the CAMDA MetaSUB consortium for analysis. Quality control of the samples was establish using FASTQC. Taxonomic profiles displaying organismal abundance of the samples were generated using MetaPhlAn2. The taxonomic profiles were split into test and training sets, pre-processed, labeled appropriately, and ran through machine learning models using Jupyter notebooks and python3 programming. Based upon prior research, the random forest machine learning model was implemented for both continent and city sample classification. Best parameters for the model were determined before implementation using random and grid search. Accuracy scores for continent and city prediction were 98% and 95% respectively. Further details for the CAMDA MetaSUB challenge can be found at the following link - http://metasub.org/.

Introduction

With over half of the world's population living in urban areas, mass-transit systems like subways and buses represent some of the most shared environments in the world. While effectively transporting millions of people around urban areas, these transit systems also serve as a home to a rich community of microbes that share DNA and RNA with their human passengers. These unseen genetic interactions play important roles in public health and disease outbreak, yet little is known about mass-transit biomes. Fortunately, metagenomics, defined as the direct analysis of genetic material found in an environment, offers an opportunity to better understand the microbial communities present in our mass-transit systems.

To analyze the genetic interactions and community compositions of mass-transit biomes, The MetaSUB International Consortium has collected whole-genome sequencing (WGS) data from 16 cities across the world. Further, the Conference on

Critical Assessment of Massive Data Analysis (CAMDA) has released their Metagenomics Forensic Challenge - predict the geographic origin of a metagenomic sample when no reference samples from that location are known. For this challenge, CAMDA has provided access to the 2019 MetaSUB WGS data. Previous work has shown that the taxonomic information derived from this data can be used as features in machine learning models to predict an unlabeled sample's city of origin. Thus, we hypothesize that geographic locations in soil and transit biomes will be distinguishable given taxonomic composition and abundance profiles. Further, we assert that the classification of "unknown" samples with no reference samples is possible given additional geographic inputs.

In this study, we plan to build off the work of the MetaSUB global consortium and utilize the data provided by CAMDA to present a metagenomic analysis of mass-transit biomes from sixteen cities across the world located within 6 continents. The goal of our study is to effectively classify whole-genome sequencing (WGS) samples according to their city and continent of origin. We hope to build a classification model to predict the origin of "unknown" metagenomic samples with high accuracy rates. A list of the provided cities with their respective abbreviations and continent locations in alphabetical order can be seen below.

City	City names	Country	Continent
AKL	Auckland	New.Zealand	Australia
BER	Berlin	Germany	Europe
BOG	Bogota	Columbia	South America
HAM	Hamilton	New Zealand	Australia
HGK	Hong Kong	China	Asia
ILR	Ilorin	Nigeria	Africa
LON	London	U.K.	Europe
MAR	Marsille	France	Europe
NYC	New York	U.S.A.	North America
OFA	Offa	Nigeria	Africa
PXO	Porto	Portugal	Europe
SAC	Sacramento	U.S.A.	North America
SAO	Sao Paulo	Brasil	South America
SOF	Sofia	Bulgaria	Europe
STO	Stockholm	Sweden	Europe
TOK	Tokyo	Japan	Asia

Table 1. List of cities used in the MetaSUB challenge

Methods

For our analysis, we used the provided WGS samples collected from 16 major cities worldwide by the MetaSUB consortium. Provided samples were first tested for quality control using FASTQC with all samples showing displaying acceptable results.

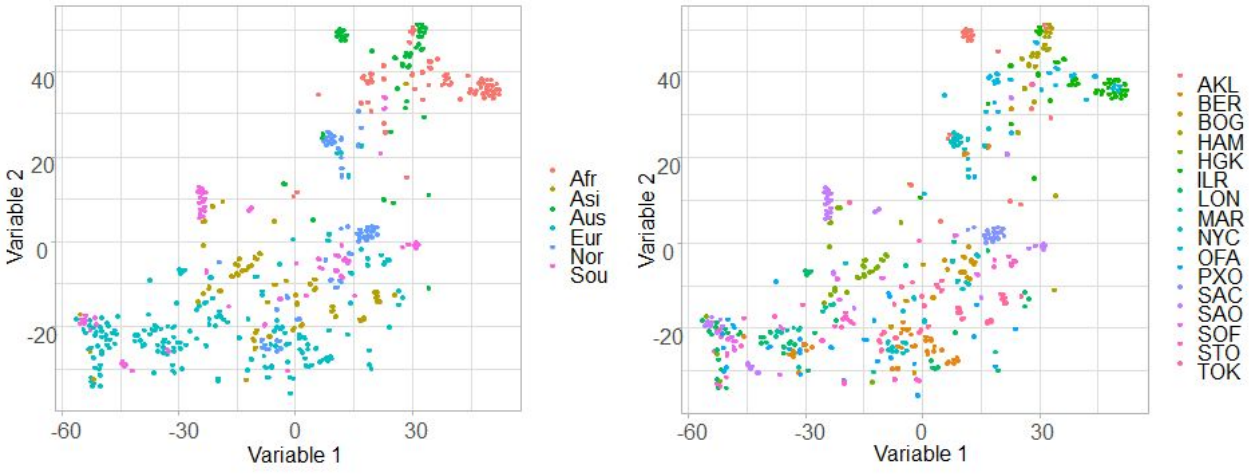
We then used MetaPhlAn2 to generate operational taxonomic unit (OTU) tables. We hypothesized that the taxonomic profiles will associate with continents/cities of origin; however, we also expect the OTU tables to be different based on the sequencing strategy used.

To first visualize the data for clusters, our OTU tables were loaded into Rstudio and a clustering plotting tool of the T-Distributed Stochastic Neighbor Embedding (TSNE) type was implemented to display clusters. The detection of obvious clusters then led us to build a machine learning algorithm that would accurately classify samples based upon sample similarity and mathematical clustering. Based upon the information provided to us by the individuals who conducted last year's MetaSUB challenge, the random forest machine learning model was therefore used due to its remarkably high accuracy scores.

Using a Jupyter notebook and python3 programming, the data was loaded into our notebook and labeled appropriately. The data set was then separated using a stratified sampling approach into training (80%) and test (20%) sets. The best parameters for our random forest model were then generated by random search and grid search found within the "sklearn" python module for both continent and city analysis. Random forest models were then implemented for the continent and city analysis, their results displayed, and corresponding confusion matrices generated as well.

Results

Visualization of the provided samples via dimensionality reduction using TSNE visualization did indeed display clusters for both the continents and cities(Figure 1). The figures result showed that most continents were well clustered, with Europe, South America and Asia not being as strongly or densely clustered. Likewise with city clustering, some cities are much more well clustered than others (SAC and ILR are especially well clustered). Overall, continents showed better clustering results as expected.



A B

Figure 1. TSNE visualization of the sample data. A shows the sample clusters in different continents (labeled in different colors), X axis is variable 1 of TSNE, and Y axis is the variable 2 of TSNE; B shows the sample clusters in different cities (labeled in different colors), X axis is variable 1 of TSNE, and Y axis is the variable 2 of TSNE. RTSNE visualization analysis of the data in R.studio: 1. Unique the sample to exclude duplicate samples; 2. RTSNE : tsne_out_p20_pca <- Rtsne(X_unique, pca=TRUE, perplexity=24, theta=0), perplexity may need to adjust; 3. Plot generated using package "ggplot2".

Identification of the best hyperparameters using both random and grid search and implementation of the random forest model yielded an accuracy score of nearly 98% for continent classification (Figure 2). Computational runtimes for both random and grid search were around 12 minutes for each. Actual implementation of the model took roughly .0177 seconds (using a PC with a 1.8 GHz Intel Core i7) for continent analysis as seen below in Figure 2.

	precision	recall	f1-score	support
Africa	1.00	1.00	1.00	25
Asia	0.96	1.00	0.98	27
Australia	1.00	1.00	1.00	18
Europe	0.94	1.00	0.97	51
North America	1.00	1.00	1.00	24
South America	1.00	0.88	0.94	33
micro avg	0.98	0.98	0.98	178
macro avg	0.98	0.98	0.98	178
weighted avg	0.98	0.98	0.98	178

```
Overall accuracy score = 97.752809
Runtime:  0.017731859999912558  seconds
```

Figure 2. Classification table for Continent Analysis. Random search parameters: {'n_estimators': 100, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 40, 'bootstrap': False}. Grid search parameters: {'bootstrap': False,

'max_depth': 100, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 4, 'n_estimators': 100}

To better understand where our random forest model incorrectly classified a sample's continent location, a confusion matrix was generated (Figure 3). The Matrix also shows the number of samples corresponding to each continent with Europe having the most by far. Europe was also the hardest continent to classify according to the matrix and was mistaken for South America three times.

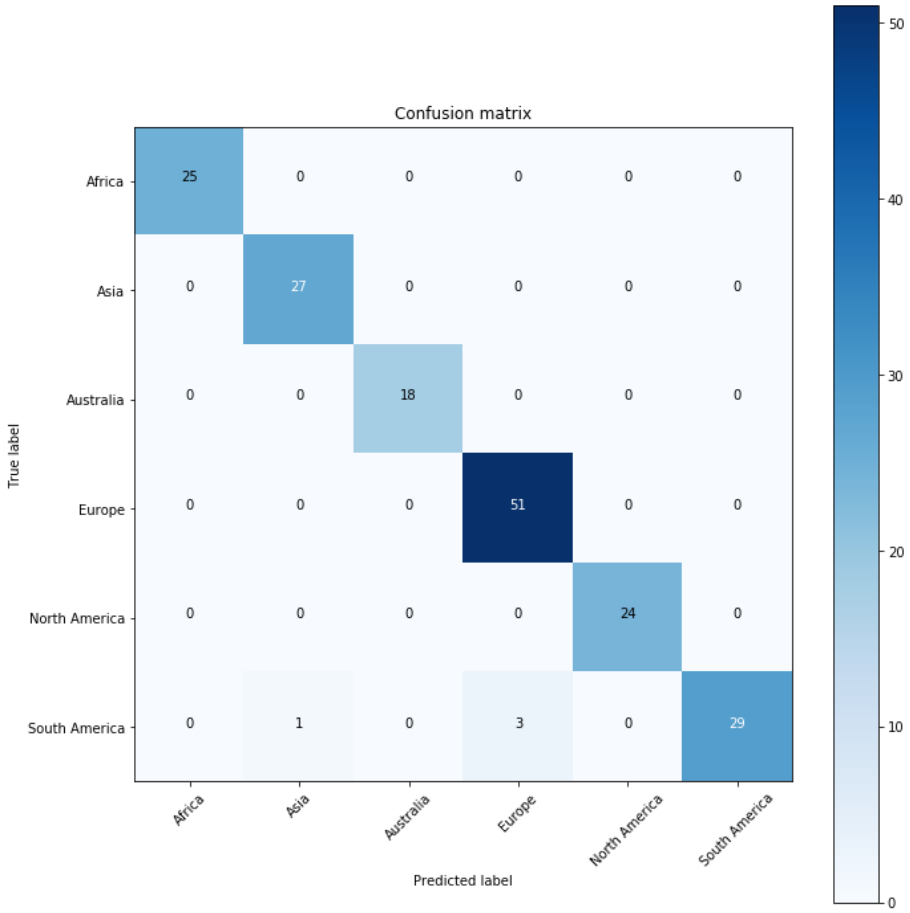


Figure 3. Confusion matrix for Continent Analysis.

As with continental classification, identification of the best hyperparameters for city classification using both random and grid search was used. Implementation of the random forest model yielded an accuracy score of 95% (Figure 4) using the identified parameters. Computational runtimes for both random and grid search were around 18 minutes for each of them. Actual implementation of the model took roughly .0377 seconds (using a PC with a 1.8 GHz Intel Core i7) for city classification analysis.

	precision	recall	f1-score	support
AKL	0.82	0.82	0.82	11
BER	0.86	1.00	0.92	12
BOG	1.00	1.00	1.00	7
HAM	1.00	0.82	0.90	11
HGK	1.00	0.80	0.89	10
ILR	1.00	1.00	1.00	20
LON	0.80	1.00	0.89	12
MAR	1.00	1.00	1.00	3
NYC	1.00	1.00	1.00	20
OFA	1.00	1.00	1.00	9
PXO	1.00	0.75	0.86	12
SAC	1.00	1.00	1.00	5
SAO	1.00	1.00	1.00	15
SOF	1.00	1.00	1.00	6
STO	0.80	1.00	0.89	8
TOK	1.00	1.00	1.00	17
micro avg	0.95	0.95	0.95	178
macro avg	0.95	0.95	0.95	178
weighted avg	0.96	0.95	0.95	178

Overall accuracy score = 94.943820
Runtime: 0.0377605229987239 seconds

Figure 4. Classification table for City Analysis. Random search parameters: {'n_estimators': 50, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': 60, 'bootstrap': False}. Grid search parameters: {'bootstrap': False, 'max_depth': 15, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 300}

To show were the random forest model made incorrect city classifications, a confusion matrix was generated (Figure 5). The number of samples corresponding to each city can be seen within the matrix as well. Overall, the European cities were much more likely to be misclassified as seen by the confusion matrix. One can also note that the European cities, if misclassified, were much more likely to be classified to another European city.

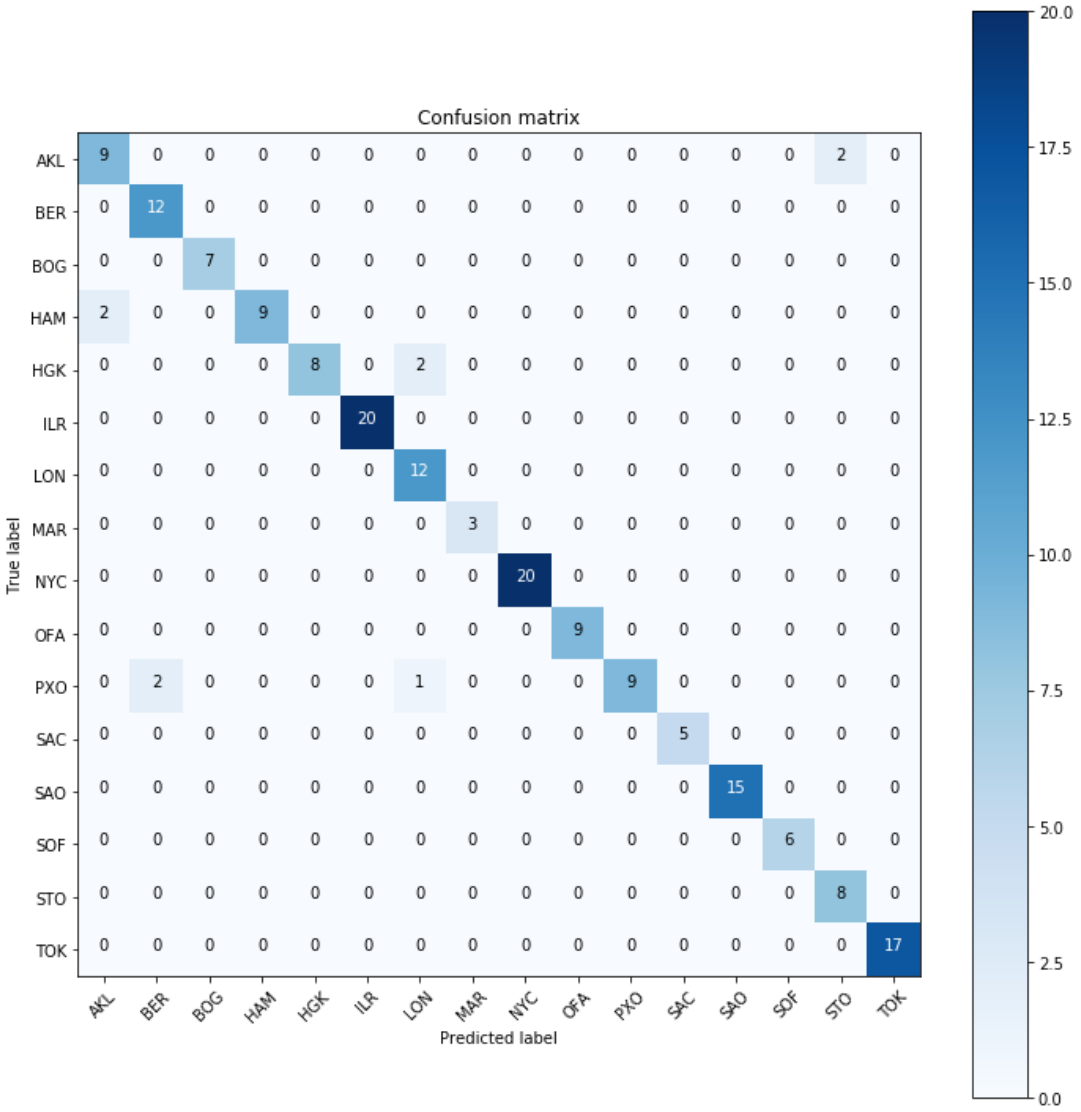


Figure 5. Confusion matrix for Continent Analysis.

Discussion

According to the TSNE visualization results, at the continent level, prediction accuracy of samples from Europe, South America and Asia will be less than the accuracy of samples from North America, Africa and Australia. These findings from our visualization of the data were also reinforced by the confusion matrices of the random forest models displayed in the results section.

Despite some cities and continents not displaying clear clusters to the naked eye, our tuning or identification of the best hyperparameters using random and grid search was very worthwhile. Implementing random forest models around these best parameters was more than able to give accurate classifications of samples. Ultimately,

the random forest models yielded accuracy scores of 98% and 95% for continent and city classification respectively.

The continent of Europe was by far the hardest to classify along with the various European cities being the hardest to properly classify as well. This is perhaps due to the large number of samples received from European cities along with other factors such as high prevalence and use of public transportation throughout Europe (leading to more diverse genetic samples). The close proximity of European cities to one another when compared to the other continents might also make for harder classification.

All in all, we are very satisfied with our accuracy scores of 98% and 95% for continent and city classification. Future work for our project could include a variety of potential steps. Implementing more machine learning algorithms, specifically deep neural networks and convolutional neural networks should display relatively high accuracy scores based upon prior research. However, the next best step to take appears to be feature selection for our taxonomic profiles. Eliminating non-descriptive features from our OTU tables should allow for increased accuracy and faster runtimes. Using 16S rRNA sequencing samples as opposed to whole genome samples for use and cross-validation of our results should also yield better models and results.

References

1. Hsu, T. *et al.* Urban Transit System Microbial Communities Differ by Surface Type and Interaction with Humans and the Environment. *mSystems* **1**, e00018-16 (2016).
2. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* 359, 320 (2018).
3. Khodakova, Anastasia S., et al. "Random whole metagenomic sequencing for forensic discrimination of soils." PloS one 9.8 (2014): e104996.
4. Yoon, Seok-Hwan, et al. "Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies." International journal of systematic and evolutionary microbiology 67.5 (2017): 1613.
5. Truong, Duy Tin, et al. "MetaPhlAn2 for enhanced metagenomic taxonomic profiling." Nature methods 12.10 (2015): 902.
6. Bushnell, Brian. BBMap: a fast, accurate, splice-aware aligner. No. LBNL-7065E. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.
7. van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). "Visualizing Data Using t-SNE" . Journal of Machine Learning Research.
8. Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.